International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
International Conference and Workshop on Communication, Computing and Virtualization (ICWCCV 2015) – www.ijais.org

Solving Big Data Problem using Hadoop File System (HDFS)

Smita Chaturvedi
ME Pursuing
TCET, Kandivali East,
Mumbai

Nivedita bhirud
ME Computer Engineering
TCET, Kandivali East,
Mumbai

Fiona Lowden
ME Computer Engineering
TCET, Kandivali East,
Mumbai

ABSTRACT
The data which is useful not only for one person but for all, that data is called as Big data or It's a data to be too big to be processed in a single machine is known as Big data. Big data are the data which are extremely large in size that may be analyses computationally to disclose the patterns, associations and trends etc. For Example: Many users visited the amazon site; in particular page for how many user visit that page, from which IP address they visit the page, for how many hours they visit the page etc information stored in the amazon site is known as the example of Big data. Huge amount of data is created by phone data, online stores and by research data. Potentially data is created fast, the data coming from different sources in various formats and not most data are worthless but some data does has low value. Hadoop solves the Big data problem using the concept HDFS (Hadoop Distributed File System). In this paper the running of map reduce code in apache Hadoop is shown. Hadoop solves the problem of Big data by storing the data in distributed form in different machines. There are plenty of data but that data have to be store in a cost effective way and process it efficiently.

Keywords
Big data, mapreduce, 3V, Eco System, HDFS, Hadoop.

1. INTRODUCTION
Big data are the data which are huge in amount that data is to be processed and managed efficiently. Big data cannot beprocessed in a single machine. For that purpose multiple machines are required in a distributed manner to process and organize those data. Organizations are creating the data from way back. As the times going on more and more data are generated by the organizations. For example: Take example of your cell phone, whenever your cell phone is on it connect to one tower and whenever the user move around the tower gets changed as per the location of the user. This information is collected by the cell phone company. The information which are collected by the company is by the nearest base station of the user. They can even trace user when he/she makes an emergency call and don't give exact location. The most data are created by the phone data, online stores, medical data and research oriented data. All these data has to be stored and processed.

First of all it has to decide that the data is big data or not for this purpose the big data is to be listed properly. Or the definition says, the data which is useful not only for one user but for all is known as Big data. For only one user data can be stored into the single spreadsheets or in a traditional database. For example: From all this four example only two are big data like; Order details for a store, All orders across hundreds of stores, A person's stock portfolio, All stock transactions for some organizations. In all these four, second and fourth are the examples of big data. The big data are the data which are in the sets of several terabytes (TB) or the data is too big to be processed by a single machine.

A) Challenges of Big data:
The biggest challenge of big data is the size. But apart from the size some challenges of big data are; most data are worthless but some data does have low value, potentially data is created very fast, the data coming from the different sources in various formats.

2. BACKGROUND
The big data stands for 3V (Volume, Variety, and Velocity). In further point it is explained in detail.

2.1 Volume:
Volume is nothing but the size of the data. In 1980 the $100,000+ for Price/ GB and in 2013 it decrease to $0.10. For storing the personal computer data reliable storages are required and for Reliable Storage, Storage Area Network (SAN) is used but it is very expensive. So the data which is very critical those kind of data is store into the SAN. For eg. Sales data can be stored but if the data is for storage but not log files. What exactly the system need is cheaper way to store the reliable data. Storing the TB of data in SAN is not that hard but to streaming the data from the SAN where central processing system is involved it will take much more time to process the data which can slower the process of retrieving the data from the SAN.

Are Data worth Storing? Before proceeding the storing of data in the storage devices, first of all the few things about the data is to be checked like is data worth storing? For eg. Transaction data like financial, government related, log file, business data, user data like videos, documents, pictures etc, sensor data like temperature, earthquake, pollution related data, medical related data like x- ray report of the patient, brain scan etc, social data like mails, twitter etc. All these are the examples of various kinds of data. The answer of worth storing of these data is yes. It is to be stored into the storage devices. All these data are to be stored and it needs to be scale the storage capacity up to massive amount. In Hadoop the data are stores in a distributed way across multiple machines.

2.2 Variety:
The data are often coming from lots of different sources in different formats. For long time the people has used databases like SQL, Mysql, or big data warehouses like in companies Oracle, IBM to store the data. The problem is to store the data in this type of system the data needs to be fit in predefined table. Most data are semi-structured or unstructured data. The

unstructured remain data arise much in different format. For eg. Bank might have list of credit card and account transactions, scans of cheques, record with customer service interactions may be recording of those phone calls. All these data are in variety of different formats and stored in traditional system. For eg. Customer service interaction number is 1-800 this data can be converted into text and if the system have recording of that phone call of actual customer's record that can be stored in the mp3 format. Later this recording can be converted into the software.

In hadoop it doesn't matter the format of data come in unlike the traditional database the data can be stored and reformat it later. To access the variety of data it is also mandatory to make a better plan to access those data. What types of data is useful to make a better plan? Sometimes most unlikely data can be extremely useful to make a better planning. For eg. Conventional system for coordinating in logistic might crosses truck to a factory to pick up a package. However might be cross truck is not the best solution perhaps is the traffic jam. So what kind of data would be helpful to making the better plan that can save money and time for the company? Current GPS, Current Plan, Traffic data, current load and fuel efficiency. All these are the important terms which are useful to make a better plan. So to make any kind of better plan lots of information's are required.

2.3 Velocity:

The speed at which the data arrives is ready to be processed is called Velocity. The user need to able to accept the data even and stored the data even if it is coming at rate of TB/day. For eg. Consider E-commerce website if user knew a product you looked in a past then the company can recommend the similar product next time you visit our site. If user spent five minutes in a particular product then the E-commerce site can sent the email to the user by informing that the particular product is on sale. If particular organization knows that the particular browser site is a first generation i-pad then they can suggest the latest model of i-pad to that particulate user. This is the huge difference to what they do before that only stores the record of purchases. But now they store and process the web server log along with the purchases record in a warehouse by which the organization can give much better shopping experience to our customer.

3. HADOOP

There are plenty of things that which can do with the big data but first thing the data needs to store in a cost effective way and then process it efficiently. It is not easy to do when the data is in mass amount. Doug is one of the founders of Hadoop. The Google published the first research paper in 2003-04 on distributed file system and about their processing framework. In the paper, what is big data, what are the challenges faced by big data and problems related to big data were mentioned. The logo of Hadoop is yellow elephant, the name Hadoop came from the name of Doug's son's toy which was yellow elephant.

3.1 Core Hadoop

The core Hadoop project consists of store in HDFS and process the data with map reduce. The key concept is to put the data up and store the collections of machine called as cluster then it can process the data in cluster. Rather than retrieving the data from the central server. The data can be retrieve it from the cluster and process it fin the cluster only. The user can add more machine as they needed more data to be store. Many people start with few machines and add as they needed. Mid-range servers are required rather than the top range server.

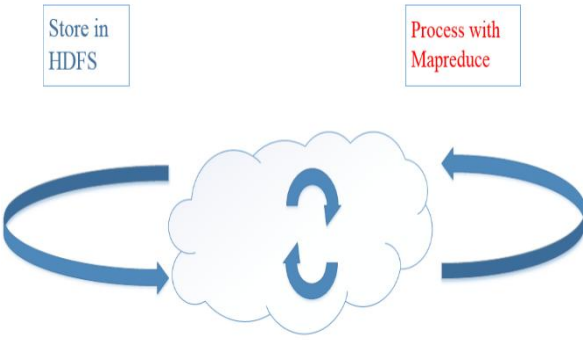

Figure 1: Hadoop

3.2 Hadoop Eco System

Basically core Hadoop is consist of HDFS and Map reduce (MR). The detail diagram includes the HDFS, MR, database and programming language. This whole system is known as Hadoop Eco System. Some software's are easier to load the data in the Hadoop system. To write the MR code the user need to know the java, python, rubi or perl etc. Lots of traditional database use SQL, Mysql query to retrieve the data. Open source project were created to access the data from the database like hive and pig. In hive the normal query like sql is written which is sown in figure. The Hive interpreter turns the sql query into MR code which can run in cluster. Pig allows writing the scripting language code rather than MR code. This code is kept into the MR and run into the cluster. If the huge amount of data is there then it will take extreme amount of time to run the code.

So new open source project implemented named impala. Impala is open source project which is used to query the data with sql which directly accesses the data from HDFS. Impala query runs very quickly as compare to hive. Hive is a long batch processing system. Other project used by many people called as Sqoop. Sqoop takes data from the traditional database system. eg Microsoft sqlserver …sqoop stored the data in a limited files so that it can be processed in the cluster with another data.

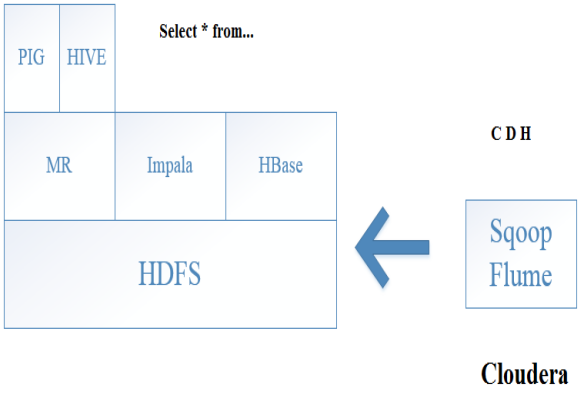

Figure 2: Hadoop Eco System

MB file is stored into the cluster that is to be divided into the chunks.

3.3 HDFS and Map reduce

Files are stored in the something called as Hadoop, Distributed, files, and system called as HDFS. It is something as normal file systems. For eg. When 150 called as Blocks. As shown in figure there are three blocks in different color as red, blue and green named blk1, blk2 and blk3 respectively. The default size for each block is 64 MB. Each block is given a unique name to identify the individual blocks in the network. The total size of data is 150 Mb but the block size is 64 MB that's why data is divided into different blocks. The addition of 64 plus 64 plus 22 is 150 MB. The file is stored in the HDFS. Each file are stored in the different nodes. In each machine one daemon is running called as a data node. One name node is there called as metadata which stores the information about the data node.

There can be multiple problems which can occur in the network:

As shown in the figure:

- Network failure in network
- Disk failure on Data Node
- Not all Data Node are used
- Block sizes are different. Eg. Some block are of size 64MB and some are 22MB
- Disk failure in name node

In above listed problems in which some can create the problems like network failure in network, disk failure on DN, disk failures on Name node. To avoid this kind of problems the Hadoop provides the solutions of data redundancy. Same are replicated at three places so that if network fails, disk fails then from the replicated copy the data can be recover.

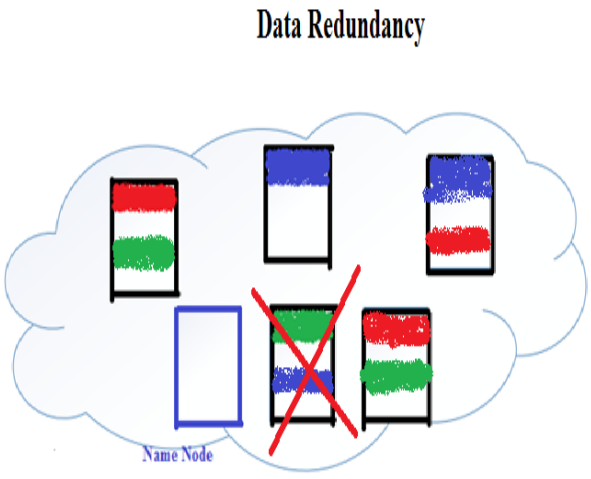

Figure 3: Data node fails

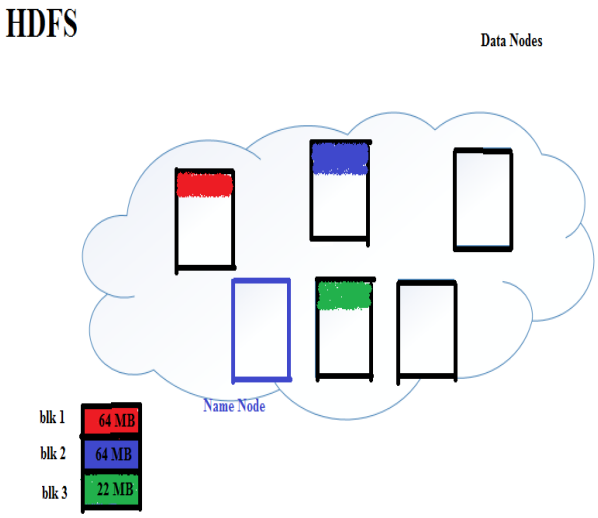

Figure 4: Distributed Data in Network

Even if any data node fails then also the same data can be recover from another data node. As shown in figure above if one data node fails same data is replicated at another node also then that data can be recover from another node.

What if the name node has hardware failure?

- Data inaccessible- If network problem is there then for the temporary time data will not be available.
- Data lost forever-if disk fails then data loss forever
- No problem

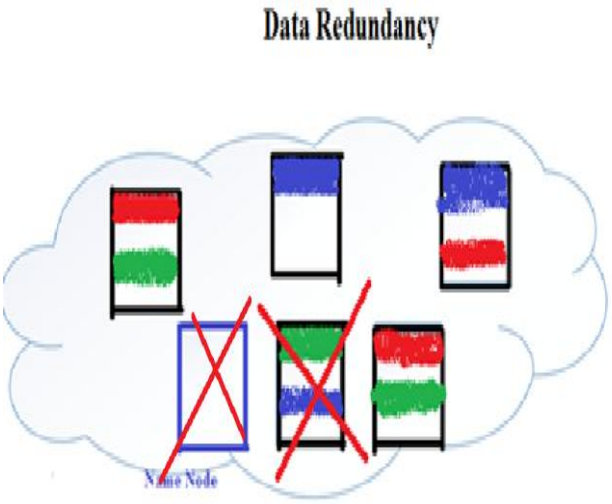

Figure 5: Data Redundancy

If name node fails then the entire data will loss. To make those data available one more node is created. It means for same data the two nodes will be there. If one node fails then the data can be recover from another node. At a time only one node will be active and other will be stand by.

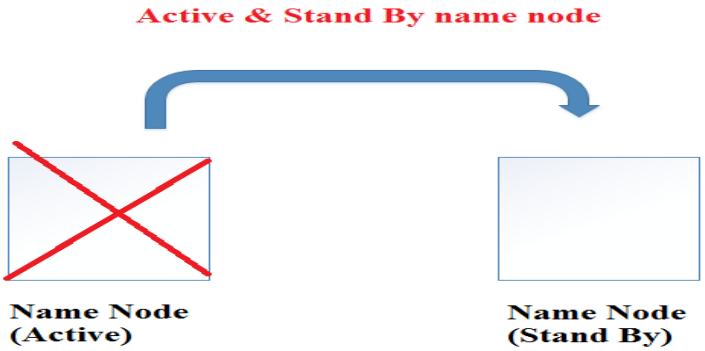

Figure 6: Active & Stand by Name Node

3.4 Map reduce

The data is stored in HDFS and processed by Map reduce. In map reduce the data is processed in distributed manner. For eg: In retailers, hash tables are used to solve the traditional system. Hash table means key with value. If data is 1 TB then what type of problem can occur? As the data is huge in amount it will take long time to run and it can be possible that it will run out of memory. So to solve this type of problem map reduce are used. MR will solve this problem more efficiently. In MR the whole work will be divided into many people.

In short, Mappers are low programs which have slow amount of data runs in parallel.For eg how to find total sales/ store then key with value is required. If store name will be the key then the value will be cost.Map reduce is the combination of two terms viz. Mapper and Reducer, Mappers will produce the intermediate records with key and value. The output of mappers will be input for reducer. By taking the help of key and value the reducer will present the final result.What happened between mapper and reducer? Shuffle and sort. The reducer keeps all the track of the data. For eg: Previous cost of the store, current cost, total sales per store, previous store, current store.After done with the mapper and reducer programs the Hadoop is used to streaming the data.

4. RESULTS AND DISCUSSIONS

The map reduce programs are written and kept in the directory to check the jobs done by Hadoop. Following are the step wise descriptions to write and execute the map reduce programs.

Step 1: Creating directory: As shown in figure new directory named inputfile is made in data directory.

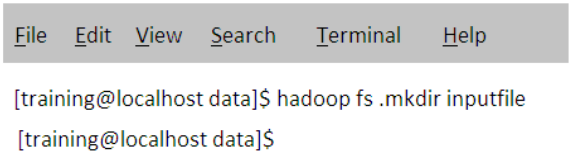

Figure 7: Creating Directory

Step 2: Putting purchase file in inputfile directory.

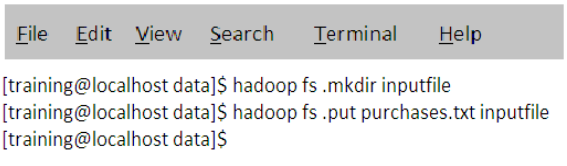

Figure 8: Placing file in Directory

Step 3: Showing Directory

As shown in the figure the highlighted part shows the name of the directory as inputfile is present in the data directory.

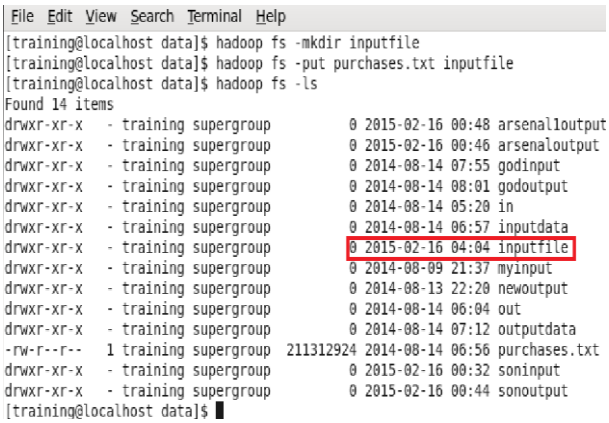

Figure 9: Showing Directory

Step 4: Content of input file:by firing the command like ls on inputfile it is showing that one file is found named purchase.txt. As in previous step no. 2 the file purchase. Txt is kept in the inputfile.

```
File  Edit  View  Search  Terminal  Help
drwxr-xr-x   - training supergroup          0 2014-08-14 06:57 inputdata
drwxr-xr-x   - training supergroup          0 2015-02-16 04:04 inputfile
drwxr-xr-x   - training supergroup          0 2014-08-09 21:37 myinput
drwxr-xr-x   - training supergroup          0 2014-08-13 22:20 newoutput
drwxr-xr-x   - training supergroup          0 2014-08-14 06:04 out
drwxr-xr-x   - training supergroup          0 2014-08-14 07:12 outputdata
-rw-r--r--   1 training supergroup  211312924 2014-08-14 06:56 purchases.txt
drwxr-xr-x   - training supergroup          0 2015-02-16 00:32 soninput
drwxr-xr-x   - training supergroup          0 2015-02-16 00:44 sonoutput
[training@localhost data]$ hadoop fs -ls inputfile
Found 1 items
-rw-r--r--   1 training supergroup  211312924 2015-02-16 04:04 inputfile/purchas
es.txt
[training@localhost data]$ clear

[training@localhost data]$ hadoop fs -ls inputfile
Found 1 items
-rw-r--r--   1 training supergroup  211312924 2015-02-16 04:04 inputfile/purchas
es.txt
[training@localhost data]$
```

Figure 10: Content of Input File

Step 5: Running the Hadoop:

To run the Hadoop four terms are very important i.e. mapper.py file, reducer.py file, the input data file and output file. If the output file is already present then it will throw error that "File is already present" for this reason everytime to get the output new file is to be created

Step 6: Output

As shown in the figure the highlight parts shows that the file is read by the Hadoop system and in which mapping and reducing of jobs are running parallel. The job ends when both mapper and reducer will reach to 100 %. The Hadoop will track the URL also.

```
[training@localhost code]$ hs mapper.py reducer.py inputfile outputfile
packageJobJar: [mapper.py, reducer.py, /tmp/hadoop-training/hadoop-unjar4870409763028034918/] [] /tmp/streamjob7613920182204355052.jar tmpDir=null
15/02/16 04:13:06 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for the same.
15/02/16 04:13:06 WARN snappy.LoadSnappy: Snappy native library is available
15/02/16 04:13:06 INFO snappy.LoadSnappy: Snappy native library loaded
15/02/16 04:13:06 INFO mapred.FileInputFormat: Total input paths to process : 1
15/02/16 04:13:07 INFO streaming.StreamJob: getLocalDirs(): [/var/lib/hadoop-hdfs/cache/training/mapred/local]
15/02/16 04:13:07 INFO streaming.StreamJob: Running job: job_201502160348_0001
15/02/16 04:13:07 INFO streaming.StreamJob: To kill this job, run:
15/02/16 04:13:07 INFO streaming.StreamJob: UNDEF/bin/hadoop job  -Dmapred.job.tracker=0.0.0.0:8021 -kill job_201502160348_0001
15/02/16 04:13:07 INFO streaming.StreamJob: Tracking URL: http://0.0.0.0:50030/jobdetails.jsp?jobid=job_201502160348_0001
15/02/16 04:13:08 INFO streaming.StreamJob:  map 0%  reduce 0%
15/02/16 04:13:21 INFO streaming.StreamJob:  map 7%  reduce 0%
15/02/16 04:13:24 INFO streaming.StreamJob:  map 14%  reduce 0%
15/02/16 04:13:27 INFO streaming.StreamJob:  map 22%  reduce 0%
15/02/16 04:13:30 INFO streaming.StreamJob:  map 30%  reduce 0%
15/02/16 04:13:33 INFO streaming.StreamJob:  map 38%  reduce 0%
15/02/16 04:13:36 INFO streaming.StreamJob:  map 44%  reduce 0%
15/02/16 04:13:39 INFO streaming.StreamJob:  map 50%  reduce 0%
15/02/16 04:13:54 INFO streaming.StreamJob:  map 69%  reduce 17%
15/02/16 04:13:56 INFO streaming.StreamJob:  map 78%  reduce 17%
15/02/16 04:13:57 INFO streaming.StreamJob:  map 83%  reduce 25%
15/02/16 04:14:00 INFO streaming.StreamJob:  map 91%  reduce 25%
15/02/16 04:14:03 INFO streaming.StreamJob:  map 100%  reduce 25%
15/02/16 04:14:06 INFO streaming.StreamJob:  map 100%  reduce 33%
15/02/16 04:14:09 INFO streaming.StreamJob:  map 100%  reduce 69%
15/02/16 04:14:12 INFO streaming.StreamJob:  map 100%  reduce 77%
15/02/16 04:14:15 INFO streaming.StreamJob:  map 100%  reduce 84%
15/02/16 04:14:18 INFO streaming.StreamJob:  map 100%  reduce 92%
15/02/16 04:14:21 INFO streaming.StreamJob:  map 100%  reduce 100%
15/02/16 04:14:24 INFO streaming.StreamJob: Job complete: job_201502160348_0001
15/02/16 04:14:24 INFO streaming.StreamJob: Output: outputfile
[training@localhost code]$
```

Figure 11: Ouput of Hadoop

Step: 7showing output file:

Three output files are generated as success, logs and part-00000.

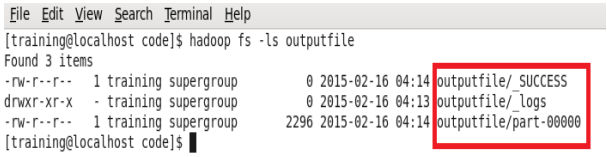

Figure 12: Showing Output File

Step 8: Creating local file:

As show in the figure the local file is created named localoutputfile.txt.

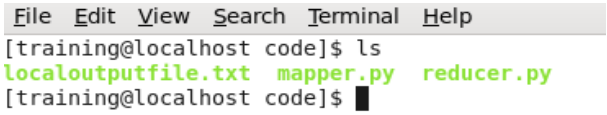

Figure 13: Creating Local File

Step 9: List local file:

The following command ls will list down all the files present in the current directory.

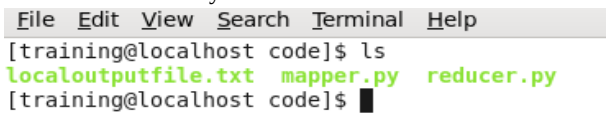

Figure 14: Listing Local File

Step 10: output

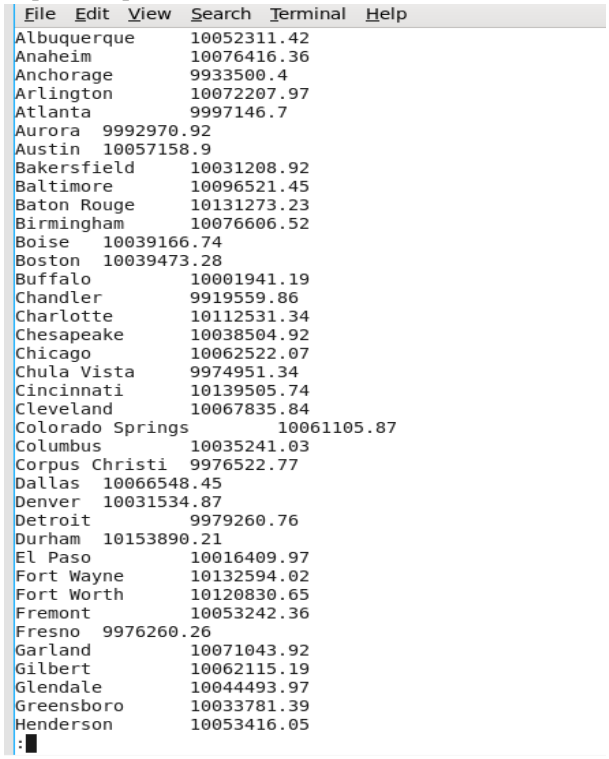

Figure 15: Final Ouput

5. CONCLUSION

As the creation of data in extremely faster way and to manage, store and process those data is handled by the Hadoop. Hadoop stores the files in HDFS (Hadoop Distributed File System) and map reduce are used to process the data. The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. Hadoop is designed to run on cheap commodity hardware, it automatically handles data replication and node failure, it does the hard work – you can focus on processing data, Cost Saving and efficient and reliable data processing. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. HDFS relaxes a few POSIX requirements to enable streaming access to file system data. HDFS was originally built as infrastructure for the Apache Nutch web search engine project. HDFS is part of the Apache Hadoop Core project.

6. ACKNOWLEDGMENTS

Our thanks to the experts Godson D'silva who have contributed towards development of this system.

7. REFERENCES

[1] The diverse and exploding digital universe. http://www.emc.com/digital universe, 2009.

[2] Hadoop. http://hadoop.apache.org, 2009.

[3] HDFS (hadoop distributed file system) architecture. http://hadoop.apache.org/common/docs/current/hdfs design.html, 2009.

[4] R. Abbott and H. Garcia-Molina. Scheduling I/O requestswith dead-lines: A performance evaluation. InProceedings of the 11th Real-TimeSystems Symposium, pages 113–124, Dec 1990.

[5] G. Candea, N. Polyzotis, and R. Vingralek. A scalable, predictable joinoperator for highly concurrent data warehouses. In35th InternationalConference on Very Large Data Bases (VLDB), 2009.

[6] The Hadoop Distributed File System : Balancing Portability, A. Hemanth, Dr. R. V. Krishniah (International Journal of Computer Engineering & Applications, Vol. III, Issue III, ISSN 2321-3469)

[7] The Data Recovery File System Systems for Hadoop Cluster- Review Paper, V.S. Karwande, Dr. S.S.Lomte, Prof. R. A. Auti (ISSN:0975-9646)

[8] The book titled with " Hadoop : The Definitive Guide" By Tom White

[9] The book titled with "Hadoop Operations" by Eric Sammer